Science & Society

EMBO reports

Could Big Data be the end of theory in science?

A few remarks on the epistemology of data-driven science

Fulvio Mazzocchi

A few years ago, Chris Anderson, former editor in chief of *Wired* magazine, published a provocative and thought-provoking article: "The end of theory: the data deluge makes the scientific method obsolete" (http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/). As the title indicates, Anderson asserted that in the era of petabyte information and supercomputing, the traditional, hypothesis-driven scientific method would become obsolete. No more theories or hypotheses, no more discussions whether the experimental results refute or support the original hypotheses. In this new era, what counts are sophisticated algorithms and statistical tools to sift through a massive amount of data to find information that could be turned into knowledge.

> *"... [an] imagined future in which the long-established way of doing scientific research is replaced by computers that divulge knowledge from data at the press of a button..."*

Anderson's essay started an intense discussion about the relative merits of data-driven research versus hypothesis-driven research that has much relevance for many areas of research, including bioinformatics, systems biology, epidemiology and ecology. Yet, his imagined future in which the long-established way of doing scientific research is replaced by computers that divulge knowledge from data at the press of a button deserves some inquiry from an epistemological point of view. Is data-driven research a genuine mode of knowledge production, or is it above all a tool to identify potentially useful information? Given the amount of scientific data available, is it now possible to dismiss the role of theoretical assumptions and hypotheses? Should this new mode of gathering information supersede the old way of doing research?

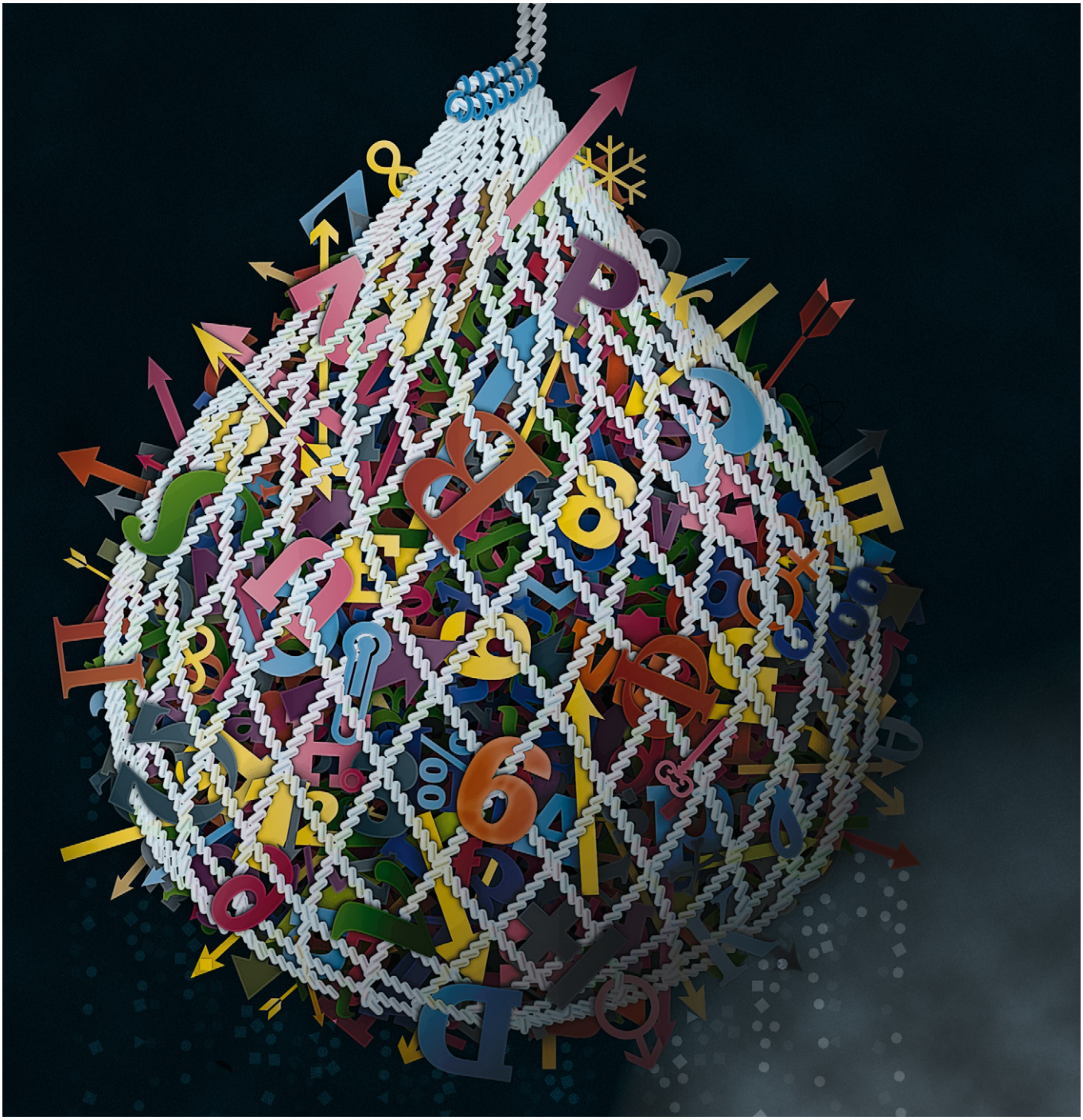
The scientific method encompasses an ongoing process of formulate a hypothesis–test with an experiment–analyze the results–reformulate the hypothesis. Such a way of proceeding has been in use for centuries and is basically accepted in our Western society as the most reliable way to produce robust knowledge.

However, Anderson is not the first to want to relegate hypotheses to a subordinate role. Francis Bacon, the "father of the scientific method" himself, in his Novum Organum (1620), argued that scientific knowledge should not be based on preconceived notions but on experimental data. Deductive reasoning, he argued, is eventually limited because setting a premise in advance of an experiment would constrain the reasoning so as to match that premise. Instead, he advocated a bottom-up approach: In contrast to deductive reasoning, which has dominated science since Aristotle, inductive reasoning should be based on facts to generalize their meaning, drawing inferences from observations and data. One of the discoveries that is frequently quoted to support this inductive approach is the laws of planetary motion by Johannes Kepler. In 1609 and 1619, Kepler, who was the assistant of Tycho Brahe, published the three laws of planetary motion based on his analysis of Brahe's observational data. These would be later verified by the laws of motion and universal gravitation in Isaac Newton's Principia. Newton was another follower of empiricism. *Hypotheses non fingo*—I frame no hypotheses—he asserted. Like Bacon, he advised a bottom-up approach, assuming the primacy of experiments, which provide empirical evidence on which to base induction.

> *"Deductive reasoning [...] is eventually limited because setting a premise in advance of an experiment would constrain the reasoning so as to match that premise."*

Big Data science renews the primacy of inductive reasoning in the form of technology-based empiricism and has inspired a view of the future in which automated data mining will lead directly to new discoveries. According to this view, the new "hypothesis-neutral" way of creating knowledge will replace traditional hypothesis-driven research. Analyzing vast volumes of data will yield novel and often surprising correlations, patterns and rules. Inasmuch as the latter emerge through a bottom-up process based on inductive processes and statistical manipulation, no theory is apparently required. Such patterns will be "born from the data" and will furnish further research

Institute for Complex Systems of the National Research Council, Monterotondo, Italy. E-mail: fulvio.mazzocchi@isc.cnr.it
DOI 10.15252/embr.201541001 | Published online 10 September 2015



hypotheses on the underlying processes, which produced the observation. In this sense, the computational approach can be seen as hypothesis generating, in contrast to the hypothesis-testing character of classical science.

According to Big Data advocates, the core of this approach is the use of inductive algorithms: "Inductive reasoning generally produces no finished status. The results of inferences are likely to alter the inferences already made. It is possible to continue the reasoning indefinitely. The best inductive algorithms can evolve: they "learn", they refine their way of processing data according to the most appropriate use which can be made [...] Permanent learning, never completed, produces an imperfect but useful knowledge. Any resemblance with the human brain is certainly not a coincidence" (http://www.paristechreview.com/2013/03/15/big-data-cartesian-thinking/).

Many valuable insights have been gained by applying this approach. In bioinformatics, for example, it has triggered a change in modeling strategies to obtain biological insights from experiments. The process of model building is driven by the massive amount of data produced and less dependent on theoretical presuppositions or hypotheses.

In the view of pioneers of DNA microarrays, "Exploration means looking around, observing, describing and mapping undiscovered territory, not testing theories or models. The goal is to discover things we neither knew nor expected, and to see relationships and connections among the elements, whether previously suspected or not. It follows that this process is not driven by hypothesis and should be as model-independent as possible. We should use the unprecedented experimental opportunities that the genome sequences provide to take a fresh, comprehensive and openminded look at every question in biology. If we succeed, we can expect that many of the new models that emerge will defy conventional wisdom" [1]. The same approach is applicable to genetic and molecular studies as well as ecosystems. The use of data analysis helps researchers to cope with the astonishing complexity of these systems, especially when large spatial and temporal scales are involved.

"*The goal is to discover things we neither knew nor expected, and to see relationships and connections among the elements, whether previously suspected or not.*"

Some Big Data advocates are making sensational claims about how this approach is going to change science itself. One example is the recent book *Big Data: A Revolution That Will Transform How We Live, Work and Think* by Mayer-Schönberger and Cukierm, which discusses three key innovations. First, the unprecedented abundance of data will guarantee a higher inclusiveness to analysis. Multiple aspects of the same problem can be investigated to provide a comprehensive picture, rather than focus on random portions of it. This reduces also the concern for sampling. Second, Big Data will allow us to lessen our yearning for exactitude. Rather than seeking accurate results under controlled and simplified conditions, scientists are driven to see in the messiness of data a reflection of the complexity of nature. Measurement errors become more acceptable. Third, and most importantly, Big Data will put a strong emphasis on correlations, that is, relations "between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone" (http://www.merriam-webster.com/dictionary/correlations). Of course, correlations are already used in science as heuristic tools and often function as the starting point for further investigation. However, this claim assumes the primacy of correlations over causal explanation or, even more radically, the replacement of the latter with the former. To put it in Anderson's words: "Petabytes allow us to say: "correlation is enough". We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot [...] Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all" (http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/).

A critical analysis of these assumptions is beyond the scope of this article, but these arguments were discussed by Sabina Leonelli, a philosopher of science at Exeter University in the UK, who questioned, for example, the idea that Big Data will cause sampling to disappear as a scientific concern: "Big Data that is made available through databases for future analysis turns out to represent highly selected phenomena, materials and contributions, to the exclusion of the majority of biological work. What is worse, this selection is not the result of scientific choices, which can therefore be taken into account when analysing the data. Rather, it is the serendipitous result of social, political, economic and technical factors, which determines which data get to travel in ways that are non-transparent and hard to reconstruct by biologists at the receiving end" [2].

Instead, in this essay, I will focus on the "no theory" thesis. While I agree that Big Data is an opportunity for scientific research, I do not believe in the presumed neutrality of numbers or the thesis that correlations will become more important than causation. As already mentioned, several sciences, such as genomics and astronomy, are generating huge data sets in the range of petabytes. Data mining techniques are increasing our capacity to find relevant patterns within these huge amounts of data. Some of these patterns do not arise from linear relations. These techniques are able to uncover complex structures in high-dimensional data that were previously unknown. This is certainly a valuable task. As stated by Mayer-Schönberger and Cukierm in their book, "the correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening. And in many situations this is good enough." However, in most cases, understanding the *why* is crucial for reaching a level of knowledge that can be used with confidence for practical applications and for making reliable predictions. "What I cannot create, I cannot understand," Richard Feynman wrote on a blackboard shortly before his death. For Feynman, truly understanding something meant being able to follow and understand each single step of the process.

As mentioned above, correlations play an important role as heuristic devices. Yet, in most cases they have to be further analyzed—using models and experiments—to assign them a meaning and to distinguish between meaningful and spurious correlations. An example of the latter comes from data mining techniques in finance, which showed a strong statistical association between the annual changes in the S&P 500 stock index and butter production in Bangladesh.

"*Rather than seeking accurate results under controlled and simplified conditions, scientists are driven to see in the messiness of data a reflection of the complexity of nature.*"

The tendency to conflate the undisputed usefulness of Big Data—which is, above all, as an information tool—with its presumed ability to provide full scientific understanding, sometimes leads Big Data specialists to overstate their claims. An example is ENCODE (Encyclopedia of DNA Elements), a large scientific project to identify all "functional" DNA elements encoded in the human genome. ENCODE involves 440 scientists from 32 laboratories worldwide, each conducting 24 types of experiment on 150 cell lines. It has generated around 15 terabytes

of data and implies a lot of data mining and analysis, with a special focus on combining different data sets in order to find and evaluate patterns.

"... correlations play an important role as heuristic devices [but] have to be further analyzed [...] to assign them a meaning"

The most relevant outcome from ENCODE is the finding that most of the human genome (about 80%) could be assigned a "biochemical function," meaning that it participates in at least one biochemical event in at least one cell type. This result, which has received much attention in the press, contrasts the notion of junk DNA —that is, DNA sequences with no apparent function—which were believed to make up more than 90 percent of the human genome. But is it really true that this concept has been debunked by the ENCODE project?

One argument concerns the notion of "function" by ENCODE: "Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure)" [3]. In light of this definition, it is possible to assign function to 80 percent of the human genome. But the ENCODE definition is clearly very loose. The American biologist Michael White and his team randomly generated 1,300 DNA sequences and found that most of these can be regarded as functional along with the biochemical criteria used by ENCODE. In this frame, it is difficult to discriminate between functional and non-functional DNA: "Most DNA will look functional at the biochemical level. The inside of a cell nucleus is a chemically active place. The real puzzle is this: how does functional DNA manage to distinguish itself from the vast excess of dead transposable elements, pseudogenes, and other accumulated junk?" (http://thefinchandpea.com/2013/07/17/using-a-null-hypothesis-to-find-function-in-the-genome/).

The main aim of ENCODE is to thoroughly measure the biochemical activities of the human genome and to supply the resulting data as resources for further studies. Biochemical activities, found with the help of computation, only suggest a function—if the notion is properly defined—but they do not demonstrate by themselves that this particular region of the genome actually does "something useful for us" (http://www.huffingtonpost.com/michael-white/media-genome-science_b_1881788.html). Much more work is required to understand whether a certain part of the genome does have a biological function and how this works—and this requires, above all, smaller-scale, hypothesis-driven research.

More data do not necessarily generate more knowledge. Data by themselves are meaningless. The idea that "with enough data, the numbers speak for themselves" hardly makes sense.

The "no theory" thesis contrasts with the fact that the collection of data is not a merely empirical activity. Science does not collect data randomly. Experiments are designed and carried out within theoretical, methodological and instrumental limitations. Instruments are designed based on prior theories and knowledge, which determine what these instruments indicate with respect to the object under investigation. Research does not examine each possible manipulation that could occur, but selects what is relevant in light of a given perspective, sometimes in order to match theoretical predictions with experience.

The collider experiments in high-energy physics illustrate this selective mode of conducting research. After the discovery of the W and Z bosons in 1983, the Standard Model of elementary particles—quarks, leptons and forces—was considered as basically proven; the only particle not yet discovered was the Higgs boson. In 2011, 18 years later, scientists at CERN's Large Hadron Collider (LHC) first observed signals of a new particle that matched the predicted mass of the Higgs boson; on July 4, 2012, CERN announced that it had finally proven its existence. This discovery was only possible with the LHC, the world's largest and most powerful particle collider and the single biggest machine ever built by humans for a specific purpose: to create particle collisions energetic enough to produce a Higgs boson.

Most elementary particles, the Higgs boson included, do not leave direct traces in detectors, because they decay very quickly. To demonstrate the existence of the original particle, scientists have to measure the decay products and track their paths back to their origin. This requires cathedral-sized detectors and millions of measurements to generate enough raw data about decay products. The LHC generates up to 600 million collisions per second and produces 15 petabytes (15 million gigabytes) of data per year. Finding the traces of elementary particles requires sifting through this deluge of data to look for specific patterns. To handle this enormous task, the Worldwide LHC Computing Grid (WLCG) that links hundreds of data processing centers around the world was created in 2002. The performance of the Grid is essential for supporting LCH experiments and releasing results quickly. Big Data, distributed computing and sophisticated data analysis all played a crucial role in the discovery of the Higgs boson—and perhaps in finding new "patterns," they might also generate new hypotheses in this field. But the discovery of the Higgs boson was not data-driven. The collider experiments were mostly driven by theoretical predictions: It is because scientists were attempting to confirm the Standard Model of elementary particles that the discovery of the Higgs boson—the only missing piece—could occur.

"Big Data, distributed computing and sophisticated data analysis all played a crucial role in the discovery of the Higgs boson [...] But the discovery of the Higgs boson was not data-driven."

Scientific research does not take place in a purely theoretical and rational environment of facts, experiments and numbers. It is carried out by human beings whose cognitive stance has been formed by many years of incorporating and developing cultural, social, rational, disciplinary ideas, preconceptions and values, together with practical knowledge. Scientists form their ideas and hypotheses based on specific theoretical and disciplinary backgrounds, which again are the result of decades or even centuries of history of scientific and philosophical thought. As stated by Henri Poincaré in

Science and Hypotheses, "It is often said that experiments should be made without preconceived ideas. That is impossible. Not only would it make every experiment fruitless, but even if we wished to do so, it could not be done."

> *"The data-driven approach constitutes a novel tool for scientific research. Yet this does not imply that it will supersede cognitive and methodological procedures..."*

Preconceived notions may take the form of tentative, explanatory hypotheses. According to Karl Popper in his book *The Logic of Scientific Discovery*, published in 1959, hypotheses function as conjectures to be checked and tested by means of empirical control. They determine what to look for and which data to collect. In their subtler form, hypotheses can be seen as a sort of basic mechanism on which the selection and interpretation of perceptual stimuli depend. Supporters of Big Data do not disprove the idea that even the computational approach involves the testing of certain assumptions, for example some search algorithm which is included in the data analysis program. But these assumptions do not provide an explanation of the phenomenon involved, merely defining strategies to identify relationships between sets of data [4].

However, this does not explain away the power and importance of preconceived notions and hypotheses, which influence how scientists plan experiments or simulations; how computational tools are designed; or the way to look at data to extrapolate regularities or correlation patterns: "Any statistical test or machine learning algorithm expresses a view of what a pattern or regularity is and any data has been collected for a reason based on what is considered appropriate to measure. One algorithm will find one kind of pattern and another will find something else. One data set will evidence some patterns and not others" [5]. Preconceived notions influence, of course, the way the discovered patterns are interpreted too. Thus, when supporters of a purely data-driven approach claim that "numbers speak for themselves," or that they are not *a priori* committed to any theoretical view, they are not doing science, but rather metaphysics.

Most objections to the "no hypothesis" or "no theory" thesis have generally been grounded on Karl Popper's view, according to which there is no such thing as pure induction. But, Thomas Kuhn's monograph, *Structure of Scientific Revolutions*, published as a book in 1962, can also offer insights. In illustrating the dynamics of "revolutionary" scientific discoveries, Kuhn emphasized the crucial role of "anomalies." By definition, anomalies can be perceived as such only *by contrast*. Pre-existing assumptions create expectations on how the world should function, and it is these assumptions and expectations that allow us to detect the odd things.

For such discoveries to occur, establishing that *there is something* that does not match our expectations is not enough. We have also to find out *what it is*. This process does not arise directly from data or numbers, but rather from a change in how we look at them, and it involves a reassessment of our beliefs and methodologies.

Similar to the emphasis on facts, the emphasis on numbers and data—which can be seen as collections of facts (e.g., values or measurements)—is another way to frame the notion or myth of the objectivity of scientific knowledge. It seems like an attempt to find in computational power that we have not found in human cognitive abilities.

However, data—even scientific data—are not "out there." Data have to be regarded as data, just like objects or facts have to be regarded as objects or facts. Yet this is far from being a trivial process. How many times in the history of science was an object, a fact or some data considered "real" and credited with a causal power although we know, today, that it was a scientific mistake? Take the example of the phlogiston or the ether. It is instructive to refer to Kuhn again: "In a sense that I am unable to explicate further, the proponents of competing paradigms practice their trades in different worlds. One contains constrained bodies that fall slowly, the other pendulums that repeat their motions again and again. In one, solutions are compounds, in the other mixtures. One is embedded in a flat, the other in a curved, matrix of space. Practicing in different worlds, the two groups of scientists see different things when they look from the same point in the same direction. Again, that is not to say that they can see anything they please. Both are looking at the world, and *what they look at has not changed* [italics added]. But in some areas they see different things, and they see them in different relations one to the other" (from *Structure of Scientific Revolutions*). What Kuhn calls into question here is not the existence of a world or reality as such. Rather, it is the possibility of accessing it in a neutral way. We look at the world through the lens of a particular vantage point, and the possibility to speak of—or even perceive—certain facts, data and objects depends on this vantage point.

Anderson's "end of theory" holds the merit of having stimulated an interesting debate and has been very effective as a provocation. At the same time, the way he posited the issue oversimplifies several important arguments that, in reason of their conceptual and philosophical complexity, should at least be treated with more prudence.

> *"... we need to investigate thoughtfully this new data-driven approach, the assumptions on which it is based, the values and biases it carries with it..."*

The data-driven approach constitutes a novel tool for scientific research. Yet this does not imply that it will supersede cognitive and methodological procedures, which have been refined during centuries of philosophical and scientific thought. There is no "end of theory" but only new opportunities. Framing the issue of Big Data in terms of oppositions, that is, deduction versus induction, hypothesis-driven versus data-driven or human versus machine, misses the point that both strategies are necessary and can complement each other. As others have argued, the inductive and deductive phases should be seen as an iterative cycle of knowledge acquisition. Likewise, technological devices can support researchers in generating, assessing and prioritizing their hypotheses. But this does not mean that human creativity has become a dispensable item in the scientific enterprise. Creativity is

different from mechanical calculus, and it is also different from seeing things in a conventional way. Rather, it involves exploring new ways of establishing connections, including implausible inferences. It is often from the making sense of the implausible that genuinely new perspectives and ideas arise.

> *"By definition, anomalies can be perceived as such only by contrast."*

There are many other issues to explore. Not only our instruments of investigation are changing but it seems that our view of the world and society is changing too. If this is the case, we need to investigate thoughtfully this new data-driven approach, the assumptions on which it is based, the values and biases it carries with it, as well as the possible consequences in the long term. For example, supporters of Big Data argue that datafication—to datafy something, sentiments and emotions included, means to set it in a quantified format in order to be tabulated and analyzed—represents "an essential enrichment in human comprehension" (from *Big Data: A Revolution That Will Transform How We Live, Work and Think)*. No doubt that some kind of understanding can be gained from this process. Are we sure, however, that data and quantity are all that count? Are we sure that this datafication of the world, and this viewing of reality as comprised essentially of information, will lead us into better world?

Conflict of interest

The author declares that he has no conflict of interest.

References

1. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21: 33 – 37
2. Leonelli S (2014) What difference does quantity make? On the epistemology of Big Data in biology. *Big Data Soc* 1: 1 – 11
3. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57 – 74
4. Kelley LA, Scott M (2001) On John Allen's critique of induction. *Bioessay* 23: 860 – 861
5. Hales D (2013) Lies, Damned Lies and Big Data. *Aid on the Edge of Chaos*, 1 February, http://aidontheedge.info/2013/02/01/lies-damned-lies-and-big-data/